Neural Incremental Speech Recognition Through Attention Transfer

Sashi Novitasari[1], Andros Tjandra[1], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]

[1] Nara Institute of Science and Technology (NAIST), Japan
[2] RIKEN Center for Advanced Intelligence Project (RIKEN AIP), Japan
{sashi.novitasari.si3, andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

Outline

I Background

II AT-ISR

III Experiments

IV Conclusion

I. Background

I Background

II AT-ISR

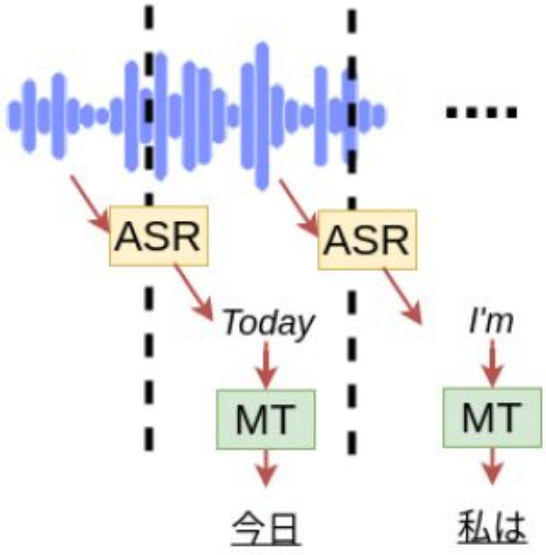
III Experiments

IV Conclusion

Simultaneous Speech Translation

- Human interpreter interprets speech in real-time
- Examples
  - Meeting
  - Lecture talk
- **Automatic** (machine)
  - Mimic human interpreter and translate incoming speech to the target language with a low delay (incremental)
  - Require ASR that can recognize speech immediately after speech start

Simultaneous speech translation system



Automatic Speech Recognition

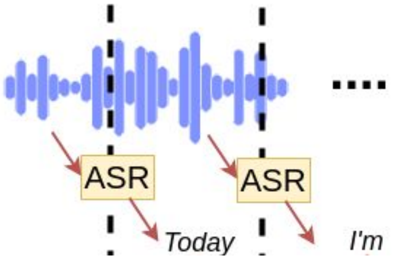
Generate transcription of a speech utterance

- **Non-incremental ASR**
  - Start recognition after speech finish
  - *State-of-the-art* ASR: Att Enc-Dec (end-to-end)

→ Not suitable for simultaneous translation

- **Incremental ASR (ISR)**
  - Recognize **without** the need for **waiting** for the **speech to finish** [Selfridge et al., 2011]
  - Part-by-part recognition challenge

→ Suitable for simultaneous translation



Incremental Speech Recognition

Current situation

- HMM-based ASR is incremental but not end-to-end [Rabiner, 1989; Gales, 2008]
- Seq2seq ISR : train by learning input-output parts alignments (e.g. Neural transducer [Jaitly et al., 2016])
- **End-to-end seq2seq ISR** → more **complex** training than standard seq2seq ASR
  - Learn the incremental step?
  - Ground alignments?
    - Alignment generation during training based on ISR model (multiple times)
    - Alignment generation using forced-alignment system (once)

Expensive

Goal

Attention Transfer Incremental Speech Recognition (AT-ISR)

AT-ISR
ISR that learns to mimic attention-based alignment from attention-based ASR

- Reliable ISR with simple construction mechanism by using attention-based seq2seq non-incremental ASR
  - ISR architecture : Att Enc-Dec ASR, identical configuration as a non-incremental ASR
  - Incremental step : Learn the attention alignment knowledge from non-incremental ASR
    □ Attention transfer
- Attention transfer : Attention knowledge transfer from teacher to student model
  - Prev. works : image recognition tasks
    - Teach another model with smaller architecture [Zaguruyko and Komodakis, 2017]
    - Domain transfer (image to video) [Li et al., 2017]
  - Has not been utilized for ISR construction

II. AT-ISR

I Background

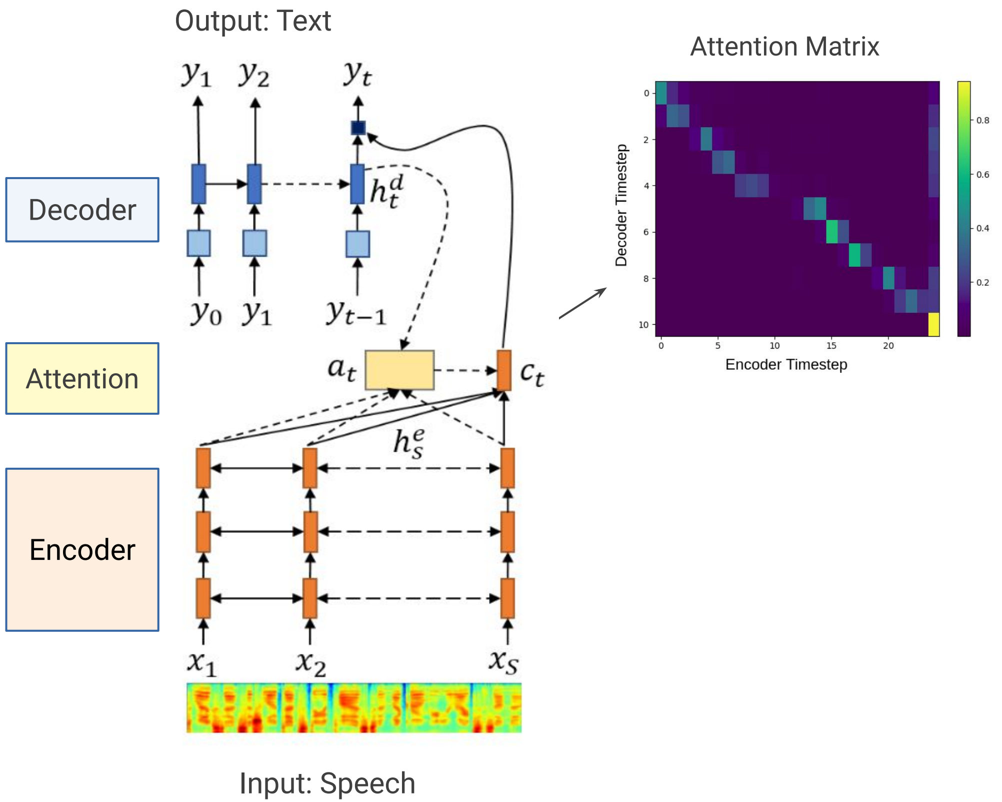
II AT-ISR

III Experiments

IV Conclusion

Overview
Sequence-to-sequence ASR

- Encoder-decoder with attention mechanism
- 3 main parts
  a. **Encoder**
     Encode input speech features ***X*** into encoder hidden state h^e
  b. **Decoder**
     Predict output ***Y*** by processing previous output, decoder hidden state h^d and encoded information
  c. **Attention**
     Calculate alignment score between encoder states (input) and decoder states (output)



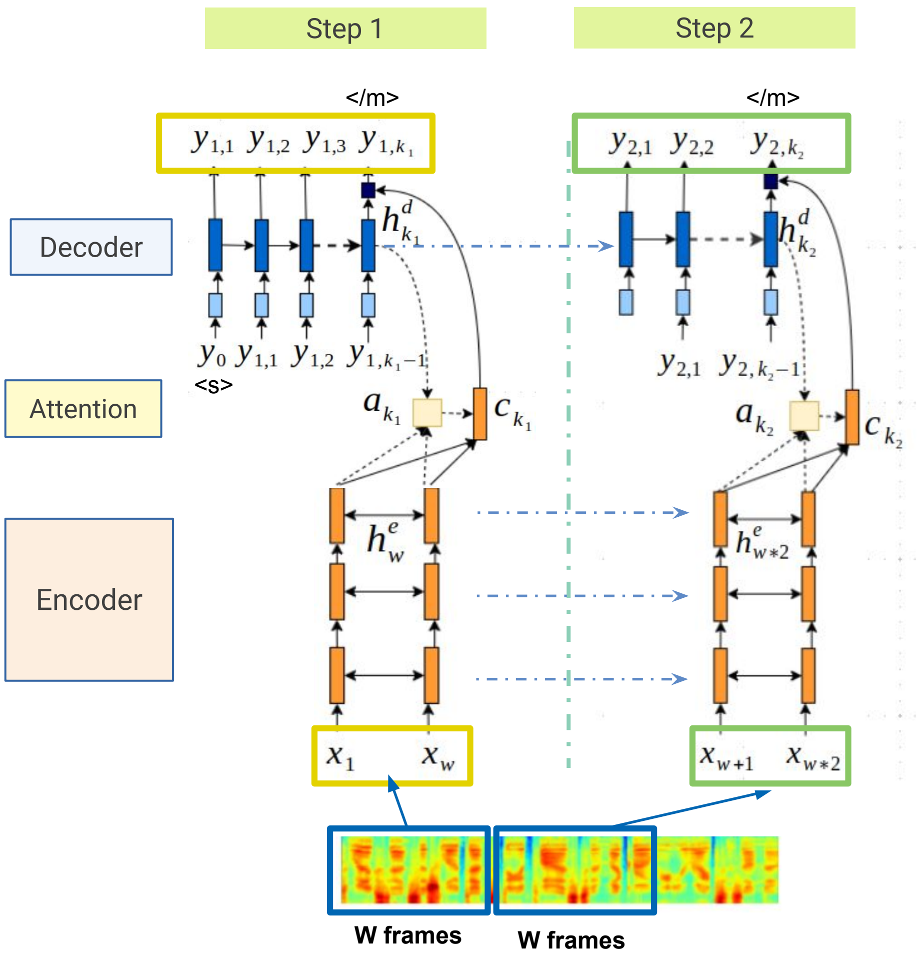
Encoder-Decoder architecture with an attention component
[Bahdanau et al, 2015], courtesy of [Tjandra et al., 2017]

AT-ISR Recognition Method

- Recognize speech **segment-by-segment sequentially**
- Delay = *W* speech frames (window)
- For each recognition step:

1. **Encode** *W* speech frames (block)
2. **Decode** for the output that aligned to the main input block, until *end-of-block </m>* token predicted or max. length reached
    - **Attend** the current input
3. **Shift** the input window *W* frames

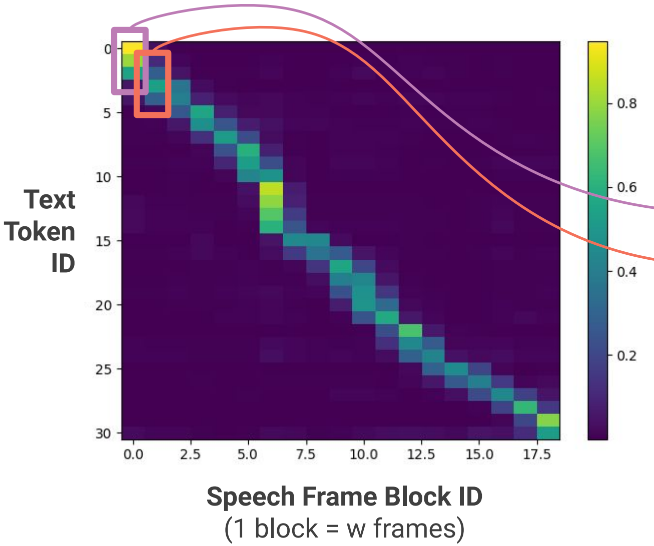
- Alignment learning → **Attention transfer**



Attention Transfer

Train ISR (student) to learn the attention-based alignment from attention-based ASR (teacher)

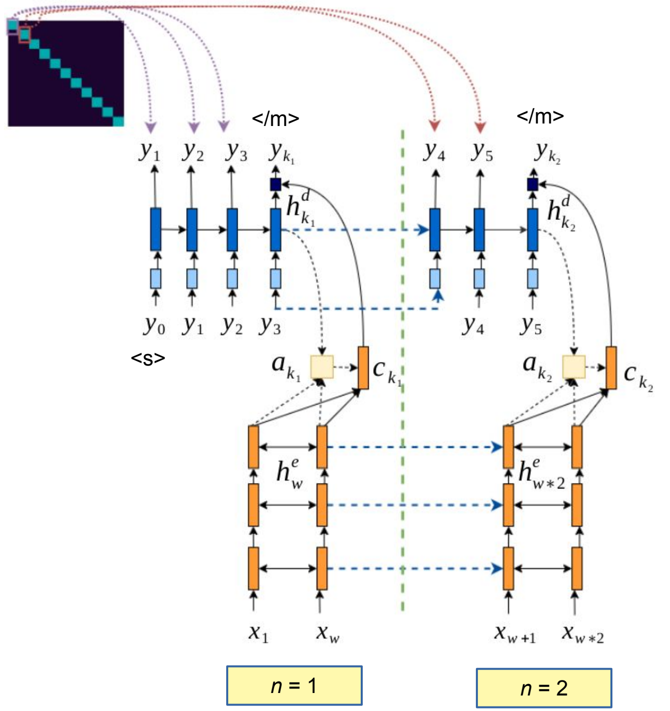
1) Extract segment-level speech-text alignment from non-incremental ASR attention matrix (alignment pair = high alignment score)



Alignment

Seg. ID (n)	Speech seg. (Xn)	Text seg. (Yn)
1	$x_1 - x_w$	$y_1 - y_3$
2	$x_{w+1} - x_{w*2}$	$y_4 - y_5$
(etc.)		

2) Train ISR by using ***Yn*** + </m> *as target of* ***Xn***



ISR delay can be managed by changing ***Xn*** and ***Yn*** size
E.g. higher delay : combine each several segments into one

III. Experiment

I Background

II AT-ISR

III **Experiment**

IV Conclusion

Experiment Setting

Data

Dataset

- LJ Speech [Ito, 2017]
  - 24 hours of speech (En)
  - Single-speaker
- Wall Street Journal (WSJ-si284) [Paul and Baker, 1992]
  - 80 hours of speech (En)
  - 280 speakers

Model Configuration

Structure: Encoder-Decoder with Attention

- Encoder
  - 1 FFN → 3 Bi-LSTM
  - Downsample 8 speech frames into 1 encoder state (ISR basic delay= 1 block = 8 frames = 0.14 sec)
- Decoder
  - Embedding → 1 LSTM
  - Output unit: character
- Same structure for non-incremental ASR (teacher) and AT-ISR (student)

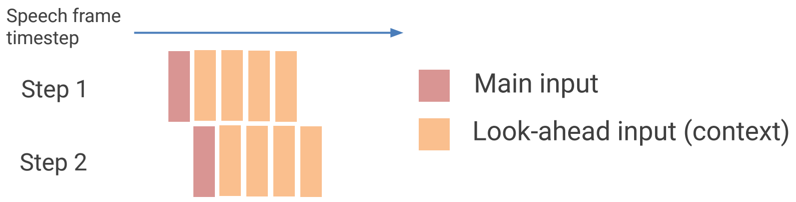
Experiment Result

Speech Recognition Result

Model	Delay (sec)	CER%
LJ Speech		
Topline ASR (Teacher)	6.54 (avg)	2.78
Baseline ISR (input/step: 1 *m*)	0.14	80.34
AT-ISR (input/step: 1 *m*)	0.14	23.04
AT-ISR (input/step: 1 *m* + 4 *la*)	0.54	4.45
WSJ-si284		
Topline ASR (Teacher)	7.88 (avg)	6.80
AT-ISR (input/step: 1 *m* + 4 *la*)	0.54	9.06

- *m* : main block (main recognition target)
- *la* : look-ahead context block (frames next to the main block)
- 1 block = 8 frames = 0.14 sec
- Baseline: Incremental recognition by using teacher ASR

ISR Input Segment



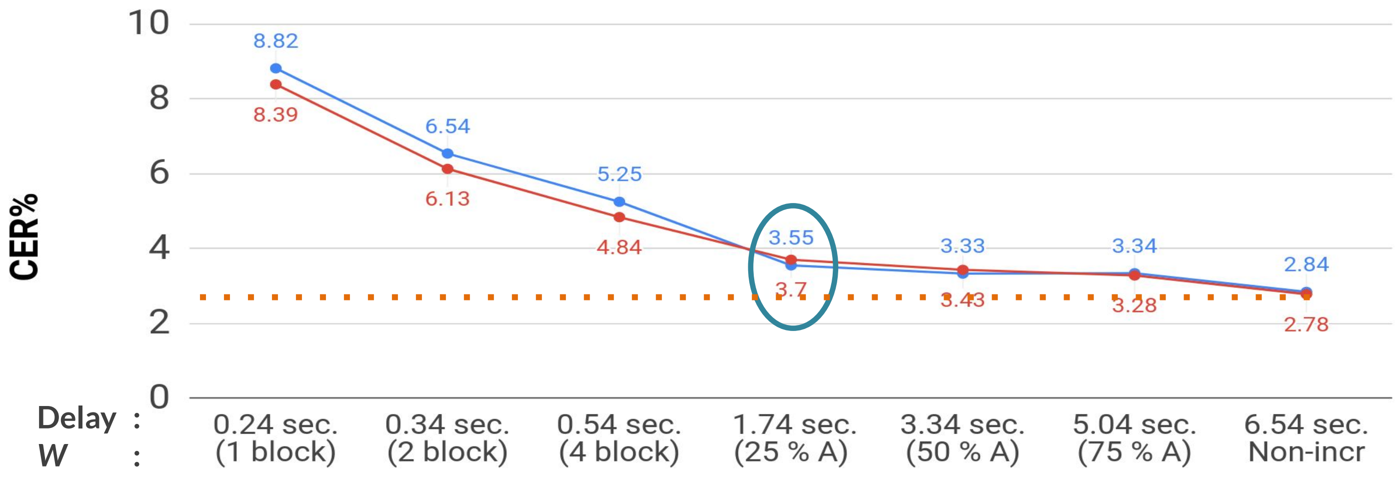
- AT-ISR has a close performance to the teacher ASR when the input includes look-ahead blocks
- Look-ahead blocks complete the information in the main block
- Non-incremental ASR (delay > 6 sec) and AT-ISR (delay = 0.54 sec) CER difference: 2%

AT-ISR performs well with a short delay by learning non-incremental ASR's knowledge

Effect of Speech Recognition Delay

- LJ Speech dataset
- Tradeoff : Higher delay □ higher performance
- Insignificant improvement after certain delay conf. □ **shortest delay with best performance**

AT-ISR CER(%) and Delay
AT-ISR Input: [*W* main + 1 look-ahead] blocks/step



A = average frame block number in an utterance in *LJ Speech* set
(6.54 sec~ 520 blocks)

1 speech block = 8 frames = ~0.14 sec

1 speech block = ~2 chars

IV. Conclusion

I Background

II AT-ISR

III Experiment

IV **Conclusion**

Conclusion

We constructed ISR that performs a low-delay speech recognition

- AT-ISR learn attention knowledge from non-incremental ASR that has an identical structure
- AT-ISR able to perform closely to the teacher by incrementally recognizing short input segments with context blocks (low latency and reliable)
  - LJ Speech CER □ teacher 2.84% (delay 6.54 sec) ; student 4.45% (delay 0.54 sec)
  - WSJ CER □ teacher 6.80% (delay 7.88 sec) ; student 9.06% (delay 0.54 sec)

Thank You

Appendix

ASR Output

Delay	ASR Output	CER%
0.14 sec (1 block)	which probably represent the burtabic condition before limbs were a quired*	11.5
0.24 sec (2 block)	which probably represent the burtabrit* condition before limbs were acquired,	7.7
1.74 sec (25% utt. length)	which probably represent the burtebrate condition before limbs were acquired,	3.8
3.34 sec (50% utt. length)	which probably represent the vertebrate condition before limbs were acquired*	1.3
6.64 sec (Non-incremental)	which probably represent the vertebrate condition before limbs were acquired,	1.3
Correct text	which probably represent the vertebrate condition before limbs were acquired;	

(*) : missing character
red character : incorrect character